Intelligent Archives in the Context of Knowledge Building Systems: Data Volume Considerations

Contract NCC5-645
Deliverable 2
File “IA-KBS data volume report v1.2”
20 April, 2005

Abstract

Algorithms for intelligent data understanding have the potential to improve the value of science data for research and applications. This potential has been demonstrated on a limited-scale basis. However, the enormous volume of science data relative to the performance of IDU algorithms raises serious questions regarding feasibility for actual use. Initial assessments of scalability indicate that it could take hundreds of years to process just one year of EOS science data. This paper takes a deeper look into the issues of data volume and scalability in an attempt to identify specific areas where intelligent data understanding algorithms could be practically applied on an operational basis to increase the utilization and overall value of collected science data. In the course of this investigation, implications for the architecture and performance of a knowledge building system are also discussed. We conclude that the actual volume that must be processed by these algorithms to produce meaningful results varies widely, from as little as 0.00006% to as much as 100% of the archive. This corresponds to data volumes ranging from 2 GB to 4 PB for current Earth science holdings. The volume to be processed by computationally intensive induction and clustering algorithms can be controlled by focusing on subsets of high level products and statistical summaries, which are probably more suitable for these algorithms anyway in terms of information content and representation. Less computationally intensive algorithms, such as matched filters for event detection, can reasonably be run against the entire archive volume. Simply unleashing complex algorithms (such as unsupervised classification) against substantial portions of the total volume is not currently feasible, and will probably remain so for at least the coming decade.

Various factors affecting the volume of data to be mined are considered in general, and also for each of the five envisioned intelligent archive capabilities (Virtual Data Products, Autonomous Event Detection, Automated Data Quality Assessment, Large Scale Data Mining, and Dynamic Feedback Loops). This provides an analytical framework for bounding the data volumes problem for various applications. An example based on fire prediction is used to illustrate the use of this framework, and reveals that the data volumes involved in Earth science make mining such data challenging but not impossible.

Contents

Preface

Purpose and Scope

This document is one of a series of papers developed under the IDU project "Intelligent Archives in the Context of Knowledge Building Systems". The purpose of this paper is to examine the feasibility of performing large scale data mining from the perspective of science data volumes (esp., Earth science) prior to examining the feasibility from other technical perspectives.

Background

Organization of This Document

The paper begins in §1.1 with a brief examination of the current and projected data volumes and computational capacity in the EOS archives to establish the scale of the overall challenge for large scale data mining. In §1.2 we examine factors that would serve to reduce (or increase) the total data volume. In §1.3 we consider the primary factors that must be considered in conjunction with data volume assessments in order to make any judgment regarding feasibility. The discussion in these sections is combined in §1.4 within the context of each envisioned capability of the intelligent archive of the future, and we summarize the results of this assessment in §1.6. Note that this document has been structured with the intent of incorporating it as an appendix into one of the other papers resulting from this project.

Revision History

Version	Date	Description of Change
0.1	09/08/04	Initial outline.
0.2	09/28/04	Work in progress for team walkthrough.
1.0	09/30/04	First draft for team review.

Approvals

Name	Signature	Date
H. K. Ramapriyan		
Ken McDonald		

Reviewers

Name	Organizations Represented
Gail McConaughy	IA-KBS Team
Chris Lynnes	IA-KBS Team
Steve Kempler	IA-KBS Team
Liping Di	IA-KBS Team
Wenli Yang	IA-KBS Team
Peisheng Zhao	IA-KBS Team

Steve Morse	IA-KBS Team

Contributors

Name	Role
David Isaac	Primary author.
IA-KBS Team	Technical direction and revisions.

1 Data Volume Considerations for IA-KBS

The preliminary results of the IA-KBS project indicate that there are a number of ways intelligent data understanding (IDU) algorithms could be applied to realize additional value from science data for research and applications [9,6]. These include a fairly detailed analysis of specific applications in agriculture [3], weather forecasting [2], and observatories [4], as well as more general uses in virtual data products [1], data quality assessment [5], and system performance optimization [8].

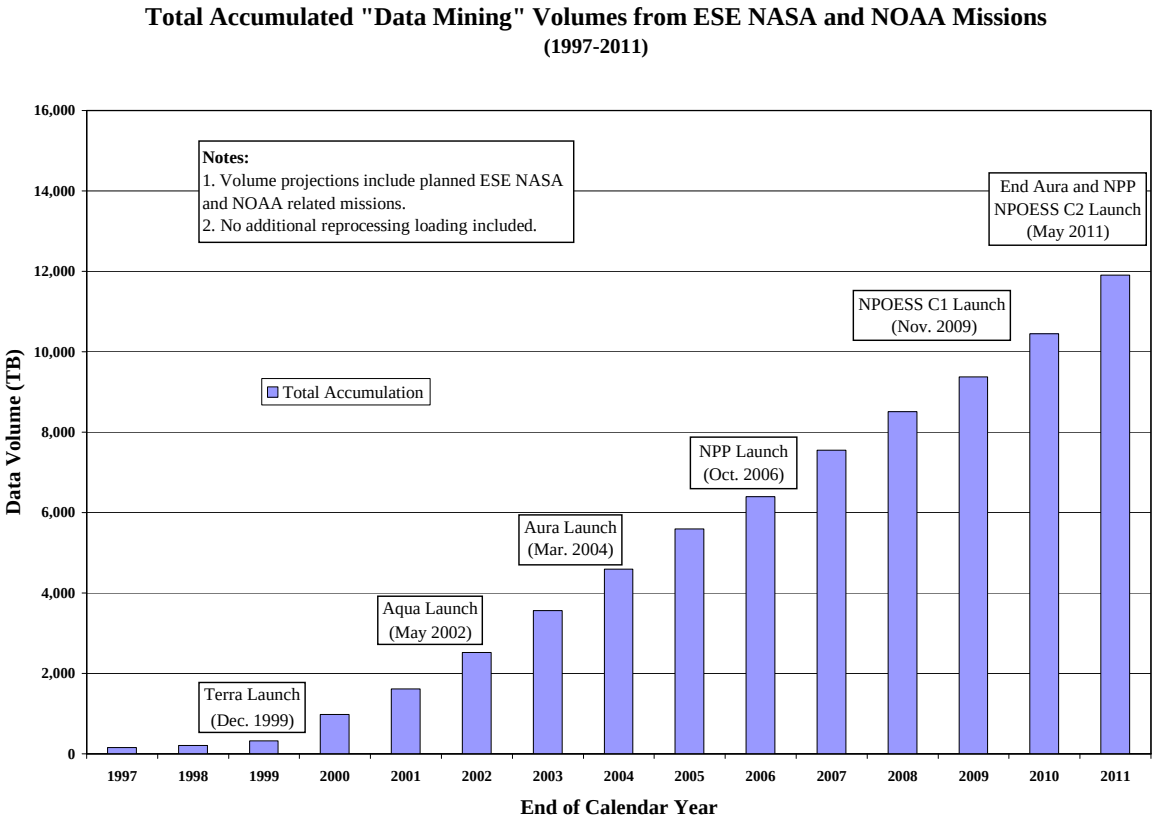
However, the enormous volume of science data relative to the performance of IDU algorithms raises serious questions regarding feasibility for actual use. Initial assessments of scalability indicate that it could take roughly two hundred years to mine just one year of EOS science data [7]. To better understand the feasibility of mining large volumes of science data in general, and Earth science data in particular, we must take a deeper look at the problem.

1.1 Scoping the Challenge

1.1.1 Total Data Volume

The total current and projected data volume of Earth science data is indeed large by any measure, exceeding four petabytes today and accumulating at roughly two terabytes per day for the foreseeable future. The figure below shows the accumulated volume of Earth science data that could be considered for mining[1].

[1] This projection excludes L0 volumes except for those missions where only L0 products are permanently archived. Raw data products equivalent to EOS Level 1A are included. EOS-ECS instrument daily archive volume requirements are taken from ESDIS F&PRS documentation. All other mission instruments also are included at 1X of the daily data rates. Missions/ platforms included in the analysis are ERBS, UARS, TOPEX / POSEIDON, RADARSAT-1, Earth Probe, OrbView-2, TRMM, Landsat 7, QuikScat, Terra, DAS, ACRIMSAT, NOAA GOES-11, CHAMP, NOAA-16, NMP/EO-1, NOAA GOES-12, Jason-1, METEOR 3, GRACE, Aqua, NOAA-17, Midori-II, ICESat-1, SORCE, Aura, TOMS Follow-on (Aura), NOAA-N, NOAA GOES-N, ISS, Cloudsat, CALIPSO, NMP/EO-3, Glory, NPP, OSTM, GCOM A1, NOAA GOES-O, GCOM B1, OCO, LDCM, NOAA-N', Aquarius, NOAA GOES-P, GPM-Core, GPM-Constellation, NPOESS C1, Hydros, NPOESS C2, LightSAR, and DSCOVR (Triana).



Most of the IDU projects tested algorithms against less than a few megabytes of data, and none appear to have tackled more than a gigabyte of data. This large disparity (nine orders of magnitude) between the volume of available for mining and the volume used in the research projects clearly raises questions as to the feasibility of moving IDU algorithms from a research setting to an operational setting.

1.1.2 Total Computing Load

A rough assessment of computing resources also provides some perspective on mining large volumes of data.

While a detailed inventory and measure of available computing resources would be time consuming to obtain, a rough approximation can be derived from a sample of current science processing. For example, MODIS daily processing requires roughly 1000 to 1700 MFLOPS to process an estimated 125 to 170 GB of data[2] for L2 and L3 data processing (respectively) [11]. This implies roughly 400 instructions per byte for contemporary science data processing.

Unfortunately, none of the IDU research projects quantified the absolute runtime performance of the algorithms used because they were primarily focused on performance in terms of the accuracy of the result. The best performance information available is that one unsupervised algorithm took 80 minutes to classify a 7000x6500 pixel Landsat scene consisting of six spectral

[2] L1 data is estimated to be 25% smaller than L2 data.

bands. Assuming 25 MFLOPS per processor and 1.5 bytes per pixel per band implies about 20,000 operations per byte.

Since science data processing systems are sized to the current workload, the above estimates imply that mining all data using an unsupervised algorithm would require 50 times the available computing resources. While this is a large difference, it is small enough that there is some hope of feasibility if we are somewhat frugal in how we apply data mining and if improvements in processing capacity continue to follow Moore's law.

Still, the total computing load is significant. A compute requirement of 400 instructions per byte on the entire data stream of 2 TB/day implies a total computing capacity of about 10 GFLOPS for current science algorithms. A data mining load 50 times greater would require 500 GFLOPS of computing capacity.

1.1.3 Total Storage Throughput

The total data volume estimate also has implications for storage throughput. Storage throughput can actually be more important that computational load because the throughput of storage devices has been increasing at a slower rate than processors. Contemporary disk drive throughput is roughly 30 MB/s. Reading an entire 4 PB archive in one month would require the equivalent throughput of only 130 disk drives operating in parallel. This somewhat surprising result would lead us to conclude that storage throughput is not a primary concern for data mining, if there was only the need to make only a single pass through the data.

The problem, of course, is that many algorithms (such as clustering and supervised classifier training) need to make more than one pass through the data. In fact, algorithms that make more than one pass through the data will commonly make hundreds or thousands of passes.[3] If the analysis performed on each pass cannot be constrained to a limited subset (say, a limited geographic area), the data must be retrieved from storage again on each pass. A thousand-fold increase in the storage throughput requirement suggested above similarly increases the level of concern about storage system throughput.

On top of this, it should be clear that one algorithm will not suffice for all purposes, so many algorithms will need to be run. A hundred algorithms potentially means a hundred passes through the data.

From this brief analysis, we can make a few observations:

- The specific type of algorithms deployed in the near term will have to be considered carefully, and in the near term global analysis may need to be restricted to certain algorithms (such as matched filters used for event detection or induction algorithms based on statistical analysis) that require only a single-pass through the data.

[3] For example, back-propagation neural network training commonly take 5000 passes to converge, though research on geospatial datasets indicates the number of iterations can be reduced to a few hundred passes with good classification performance. See, for example, Gordon German, "Neural network classifiers for GIS data: improved search strategies", http://www.geovista.psu.edu/sites/geocomp99/Gc99/093/gc_093.htm.

- In the longer term, new system architectures that allow data mining algorithms to be distributed down into the storage system could provide greater parallelism to address storage throughput concerns.
- To effectively distribute data mining, new distributed algorithms that combine local analysis with global information sharing, such as those investigated by Kumar, will be needed.
- Processing schemes that allow multiple algorithms to share a data pipeline and avoid redundant data retrievals will be needed as the number of algorithms increases.

1.2 Examining the Challenge

Based on the initial scope of the challenge, it is clear that a simplistic approach of unleashing unsupervised algorithms on the entire volume of an archive is not feasible within the next decade. The following sections take a closer look at various aspects of the problem and data mining to identify potential approaches to finding an appropriate fit.

1.2.1 Data Product Levels

The most obvious response to an overwhelming volume of data is to only mine a subset of that data.

Much of the information contained within an archive may be redundant, having been processed through a variety of levels from raw instrument readings (L0) to gridded science parameters (L3). For science purposes, the lower level data products must be archived since there is no guarantee that the information there can be totally recovered from the higher level products. By the same token, it is reasonable to assume that most interesting information is in fact preserved across data levels.

An examination of statistics for an instrument such as MODIS shows that L0, L1, L2, and L3 data products respectively account for 10%, 33%, 43%, and 14% of the total data volume [11,10]. A broader analysis of DAAC holdings shows the following distribution.

Discipline	Level 0 Pct	Level 1 Pct	Level 2&3 Pct
Land	21%	12%	66%
Atmosphere, Precip, Bio	8%	63%	29%
Radiation, Clouds, Troposphere, Chemistry	14%	83%	3%
Cryosphere	1%	0%	99%
Average	**13%**	**50%**	**37%**

Thus, there is an opportunity to reduce the total data volume requirement by roughly an order of magnitude by mining only, say, L0 or L3 data.

The choice of which level to use for data mining depends in part on the goal. L0 data and L1 data products, being earlier in the processing chain, might be the best candidates for data quality assessment so that data quality issues originating at the sensor can be identified promptly, though change detection might be difficult without a consistent geolocation for periodic samples. L1 and L2 data products are calibrated and spatially located, and thus might be good candidates for event/change detection. L3 data products are gridded and therefore more easily combined with

other data products, and might be the best candidates for scientific knowledge discovery of large-scale phenomena.

1.2.2 Earth Science Disciplines & Data Usage

Earth science data products are associated with different disciplines of study. Focusing data mining efforts on a single discipline is one way to reduce the total data volume to be analyzed. The following table shows the approximate percentage of data in the EOSDIS archives associated with different disciplines, and the percentage of archived data that is distributed to users.[4]

Discipline	Archive Pct	Distribution Pct
Land	29%	25%
Atmosphere, Precip, Bio	57%	20%
Radiation, Clouds, Troposphere, Chemistry	13%	23%
Cryosphere	1%	11%
Total	**100%**	**22%**

Obviously, any given analysis focused solely within any one of these disciplines would need to examine at most the archived percentage of the data. The meaning of the amount of data distributed is more difficult to interpret. On one hand, this data may represent that portion of the archive that is of most interest, and thus data mining could be focused within that subset. For example, the fraction not being distributed could include large volumes of lower level products that, while required to generate higher level products, are not themselves directly of interest. On the other hand, the large percentage not currently being distributed could reflect the constraints of human capacity and limited funding for scientific research. In that case, the most value from data mining might be realized by examining that data that is currently being ignored. Further analysis, including an assessment of data distributed by product level, is required to better interpret these observations.

1.2.3 Other Subsets

The level of a data product and associated discipline can be viewed as just two of many attributes that could be used as the basis of further subsetting the total volume of archived data. Other attributes include source instrument, observation time and location, data quality, and percent cloud-free. Data can also be subsetted based on science parameter or spectral band. In most cases, selecting a judicious subset of the data will improve the ability of a data mining algorithm to extract useful information. For example, cloud free images over land will be most useful for learning something about land cover, whereas cloudy images over the ocean might be useful for learning about something else. As another example, large volumes of raw SAR data might be safely excluded when mining information from radiance measurements. The point is not to leave all but a small subset of the archived data unexplored, but that *all* data need not be examined by *every* algorithm.

[4] Data volume estimates courtesy of the ESDIS project as documented in [7].

In the intelligent archive, data mining algorithms should have the ability to distinguish between relevant and irrelevant subsets of data according to their purpose. This could be as simple as a filter applied to the existing metadata.

1.2.4 Sampling and Statistical Summaries

Sampling is a set of specific methods for subsetting data. While potentially useful for reducing the total data volume, sampling must be applied with caution.

One of the hazards of sampling is that infrequent (or isolated) but highly correlated or meaningful events may be missed. In the sciences, it is not unusual for such events (including natural phenomena such as lightning strikes, volcanic eruptions, and tropical storms) to be of greatest interest. And one of the major strengths of data mining is the ability to extract useful knowledge from such events. Another hazard of sampling is that it obscures any time-periodic signal in the data that is greater than half the sampling frequency.

With these cautions in mind, we must remember that most science data already represents a sampling of the real world. The question is really not whether nor not sampling is appropriate, but what frequency of sampling is appropriate. Depending on the data mining goals, the sampling frequency available from the source instrument may be too low, just right, or too high.

In the intelligent archive, data mining algorithms should have the capability to resample data appropriate to their purpose.

1.2.5 Training Sets

In spite of the limitation of subsetting in general and sampling in particular, it is common practice to use only small subsets of data to train supervised classification algorithms. Training sets are typically a small contiguous block or random sample of the total data volume.

The proper size for a training set will vary depending on the type of knowledge or information one hopes to extract from the data. For example, is the user interested in a global or local phenomena? The training set will need to provide sufficient coverage of different conditions (e.g., geospatial variation, diurnal variation, seasonal variation, etc.) to produce a result that will be robust when applied to data containing these different conditions.

While a general statement about the size of training sets cannot be made, an illustrative example may be useful. Analysis of related to landcover might need a training set covering a few dozen types of landcover in varying proportions at a few times of the day over four seasons. A few hundred to a few thousand data points could adequately cover all these cases...the tiniest fraction of an archive. This, it is clear that for certain algorithms (e.g., supervised algorithms and rule induction algorithms) and for certain purposes, the data volume that needs to be considered is vastly less than the total archive volume.

1.2.6 Statistical Summaries

Depending on the data mining goals, the use of statistical summaries of the data may be sufficient or even preferable compared to using full-precision source data. For example, monthly 2500 km^2 average averages may be more appropriate for inducing knowledge about climate than daily 5 m^2 readings. [13] Statistical summaries directly and substantially reduce the volume of

data volume that must be examined, e.g., an effective 30x reduction for monthly summaries compared to daily measures, or an effective 40,000x reduction for 1 km resolution compared to 5 m resolution.

In most ways, statistical summaries can be treated as higher level data products, so the same points discussed under "Data Product Levels" above applies here as well.

Granule level metadata is also a type of statistical summary commonly found in Earth science archives. However, the level of summarization is probably too great for such metadata to be used directly for data mining except for inferring data quality assessment rules. Of course, metadata would likely be useful for selecting the data in the archive to be mined by any given algorithm.

1.2.7 Derived Data

Like humans, data mining algorithms have difficultly identifying associations with implicit information (e.g., that the date 10/24/04 is on a day of type "weekend" in a season called "fall"). For data mining algorithms to be effective, such implicit information must often be made explicit. The resulting derived data can increase the total data volume, sometimes substantially. For each Earth science measure, it is conceivable that there may be a dozen derived attributes relating to space and time for each measure. This represents a potential order of magnitude increase in the volume of data to be mined.

1.3 Additional Considerations

1.3.1 Algorithm Computational Complexity

Data mining algorithms vary widely in their computational complexity. In this regard, data mining algorithms can be grouped into three levels of computational complexity.

- High. This would include unsupervised classifiers (clustering algorithms), which typically examining each data point many times as they recursively explore different candidate clusters. Further, these algorithms must be run on all the data every time, since they classify instances of observations but do not produce a compact model that can be used to efficiently classify additional instances. Runtime for these algorithms can grow geometrically as either the number of observations or the number of dimensions increases.

- Medium. This includes decision trees, generalized rule induction, which generate and utilize statistical summaries of the data to build compact models of the data, which can be used to evaluate additional instances. Runtime for these algorithms can grow geometrically as the number of dimensions increases, but typically grows only linearly as the number of observations increases.

- Low. This includes simple matched filters which may be the output from a more complex algorithm. Runtime for these algorithms is typically linear with respect to both the number of observations and the number of dimensions.

Low complexity algorithms have performance characteristics similar to current science algorithms, and could conceivably be applied to the total data volume, even in real-time. These

algorithms are useful for extracting additional information from data, but do not really produce knowledge. The medium complexity algorithms directly produce knowledge in the form of an induced model that represents the data. They are typically designed to be run against representative subset of the data, and the value of running them against the total archive volume is not clear. It is probably not feasible to run high complexity algorithms against the entire data volume, although there may be some value in doing so (e.g., new phenomena could be identified by the presence of new clusters). Instead, these algorithms could be used to generate partially-labeled datasets, which are then further analyzed by humans or supervised classifiers to generate knowledge in the form of compact models representing the data.

1.3.2 Dimensions and Cardinality

Most data mining algorithms are very sensitive to both the number of dimensions (fields) in the data and the cardinality (number of possible values) of the categorical dimensions. For example, generalized rule induction algorithms will calculate statistics for various combinations of every value of all categorical fields. The resulting geometric explosion can quickly overwhelm the physical memory of even the largest computers, so calculation of the statistics quickly becomes a problem of randomly accessing large amounts of disk storage. As a specific example, brute force rule induction on just 10 fields with 10 possible values would require updating statistics for more than 10 billion combinations of values.

Thus, for the intelligent archive, scalability requires deeper consideration than just data volumes alone. In geospatial datasets, the basic number of dimensions is relatively fixed, corresponding to a fixed number of attributes about location, time, and conditions during the observation. However, the number of different science parameters or spectral bands to be considered could vary widely. For this reason, knowledge building in the intelligent archive might start simply at first, considering the relationship between only a few science measures. As computational capacity increases, a broader range of interrelationships could be considered.

1.3.3 Data Mining Output

The volume of information or knowledge output from data mining will often be inconsequential compared to the volume of mined data. For neural networks, decision trees, generalize rule induction, and similar algorithms, the output is typically a small set of equations or constants representing equation parameters. This output can be represented in a few dozen bytes. For support vector machines, the output is an equation represented by a relatively small number (dozens to thousands) of support vectors and weights. This output can be represented in a few thousand bytes.

For matched filters and clustering algorithms, however, the output may be a class identifier or class membership weight for every data point examined. If the input data consists of, say, less than ten dimensions, then the output could be more than a tenth the size of the original data. In the extreme case, where every pixel is labeled by the algorithm and the primary dimensions of the input data (e.g., location, time) are stored implicitly rather than with each pixel, the output could equal the input data size.

1.4 Feasibility of Envisioned Capabilities

A number of different capabilities are envisioned for the intelligent archive, including virtual data products, autonomous event detection, automated data quality assessment, large scale data mining, and dynamic feedback loops. Realizing these capabilities will likely require different algorithms operating on different volumes of data. As a result, the feasibility of implementing each capability, particularly in the short term, may vary considerably. The table below summarizes the potential data volume reduction that could be realized in the various areas discussed above for each envisioned capability. The aggregate data volume reduction is based on the product of reductions across all areas. Note that the percentages shown are notional and only intended to give a rough sense of the degree that the total archive volume could be reduced while still providing a meaningful level of capability. The aggregate data volume is the nominal volume to be mined assuming the estimated percentages are applied to a 4 PB archive.

	Virtual Data Products	Autonomous Event Detection	Automated Data Quality Assessment	Large Scale Data Mining	Dynamic Feedback Loops
Data Product Levels (Input)[5]	L0-L2 ~90%	L1, L2, or L3 ~10-30%	All Levels ~100%	L3 (primarily) ~10%	L0 or L1 ~10-30%
Other Subsets	Selected Time/Loc ~ 50%	Selected Products ~10%	All Data ~100%	Selected Products/Parms ~1-10%	Selected Instruments ~10%
Sampling	User Criteria ~75%	All Data ~100%	Representative ~.0001-100%	Representative ~1-100%	All Data ~100%
Statistical Summaries	User Criteria ~75%	Daily 1 km ~3%	Varies[6] ~.0001-100%	Daily 1km ~3%	Varies ~3-100%
Derived Data	None ~100%	Model Parms ~200%	Model Parms ~200%	Time/Loc ~200-1000%	Time/Loc ~200%
Aggregate Input Data Volume	25% (1 PB)	.06-.2% (2-8 TB)	.0002-100% (8 GB-8 PB)	.00006-.6% (2 GB-20 TB)	.06%-6% (2-200 TB)
Algorithm Computational Complexity	Low	Low - Med	Low	Med-High	Low
Output Data Volume (% of input volume)	Products[7] ~100%	Event Time/Loc ~0%	Data Quality Flags[7] (File/Pixel) ~.0001-100%	Statistical Model or Labeled Pixels ~0-100%	Tasking Request ~0%

Table 1.4-1 Estimated reduction of total archive data volume due to different factors for each capability envisioned for a knowledge building system.

[5] Percentages are based on MODIS products. L1 percentage within atmospheric disciplines will be higer. See section 1.2.1 "Data Product Levels".

[6] Sampling and statistical summarization percentages are not compounded since one would not typically sample a summary in this case.

[7] Note that this volume is included in, not in addition to, the archive volume.

From this table, it is clear that there is a broad range of data volumes that could be considered for data mining. The following sections discuss this range of volumes for each capability in more detail.

1.4.1 Virtual Data Products

Realizing a virtual data product capability requires the ability to dynamically generate new products from existing data. In the conceptual specimen architecture, this consists primarily of a product generation function and peer-to-peer coordination service. For input triggers needed to pre-generate products intelligently, the virtual product capability relies on the autonomous event detection capability, which in turn uses the large scale data mining capability to generate statistical filters identifying relevant events. The data volume handled by those dependent capabilities is considered separately.

The data volume that must be handled by the virtual data product capability, then, is a function of end-user requests for data. If this capability is to provide at least the same level of service as current systems, and assuming that all current data products are in fact used, the virtual data product capability will have to generate all data products at all levels (L1-L3) using predecessor levels as the source (L0-L2).

The hope is that users are interested in only a subset of measurements based on the time and location of acquisition or other factors. For example, field studies are generally only be interested in very limited temporal-spatial regions. Global L3 products, of course, will pull tremendous amounts of data through the L0-L3 processing chain. As discussed in "

Earth Science Disciplines & Data Usage" above, only 22% of archived data is actually disseminated. The amount of data used (vs. disseminated) is difficult to determine because of factors like standing data subscriptions, but it likely to be substantially smaller. On the other hand, access of higher level data products will require the generation of intermediate data products, even though those intermediate products themselves may not be distributed. So for subset considerations we guess that perhaps only half of the total volume would be referenced.

Sampling and statistical summaries could readily be employed by a virtual data product capability to determine the need to generate full resolution products. In fact this is done somewhat today, as some algorithms only process granules meeting certain criteria such as a low percentage of cloud cover.

The anticipated net reduction in data that must be handled by this capability is relatively small.[8] The algorithms, however, are also relatively low in complexity, similar or identical to today's science algorithms. Since current science algorithms operate today on the full volume of data, virtual data products are by definition feasible, at least from a data volumes standpoint. Of course, further investigation of latency given a level of computing capacity is still needed.

1.4.2 Autonomous Event Detection

Autonomous event detection involves constant monitoring of the incoming data stream or retrospective content-based search of the archive to identify events using pattern matching. It

[8] The primary "win" is that all the resulting data products need not be stored.

also relies on the large scale data mining capability to generate statistical signatures or patterns that will be used to identify events. The data volume handled by that dependent capability is considered separately.

Different events will be best identified at different levels. Most events of scientific interest are likely to be most easily identified in L3 products because change detection is easiest when one focuses on a single consistent location. However, event detection at lower levels is desirable in certain cases, such as when performing data collection and higher level processing conditionally depending on the content of the observations, or when detecting catastrophic events that require low response times. Events that are clearly identifiable outside the context of long term observations are good candidates for lower level processing. The bottom line from the perspective of data volume handling is that a relatively few L3 products should be useful for identifying a large number of events of interest, with some optimization by developing filters on lower level products possible.

Sampling will likely not be very useful, since one of the main points of autonomous event detection is to detect rare events that might otherwise be missed by end users. However, statistical summaries are likely to be very useful. We speculate that 1 km daily averages will be sufficient for identifying a large number of events of interest, including landcover changes, forest fires, severe storms, volcanic eruptions.

Because event detection relies primarily on matched filters, the only derived data required will be the (relatively few) such parameters included in the filter.

The matched filters used for event detection will have a relatively low computational complexity. Combined with a moderate total data volume, this bodes well for the feasibility of autonomous event detection. That said, detection of some events will require a substantial amount of analysis, such as that required for edge detection.[9] These may remain significant challenges well into the future.

1.4.3 Automated Data Quality Assessment

For the sake of a discussion of data volumes, two basic types of data quality assessment need to be considered. The first, detection of anomalies that indicate a data quality issue, is in operation to event detection. The second, assigning data quality labels based on various observed factors, is similar or identical to current data quality algorithms, with the exception that the quality function may be derived through large scale data mining rather than manually constructed.

Because of the importance of data quality in science data systems and the potential for introducing defects at any point in the processing chain, data quality assessment should be performed on all products at all levels. However, sampling and statistical summaries could be a viable means of reducing data volumes, since such techniques are regularly used today. In some cases, it may be sufficient to validate a few data points among millions in a given file, which would be good evidence that the software that produced it ran correctly. In other cases, it may be necessary to check for unusual deviations in every data point, such as an unexpected change overnight in land cover from an urban to dense forest classification.

[9] See, for example, [14]

Like event detection, data quality assessment must use derived data that are required parameters to any filter or other statistical model used. Also like event detection, data quality assessment will have relatively low computational complexity, without the occasional higher complexity analyses required there. This low complexity, combined with opportunities to reduce data volumes through sampling bode well for the feasibility of implementing automated data quality assessment.

1.4.4 Large Scale Data Mining

This capability performs all the heavy lifting for the intelligent archive, with its output used by all of the other capabilities as well as by end users.

For generating knowledge of interest to end users, L3 products are likely to be most useful. This is true for two reasons. First, the resulting models will be easiest to interpret when expressed in terms of recognized science parameters and stable geolocations. Second, many algorithms will only be able to operate effectively with consistently georeferenced data. For generating statistical models used by the other envisioned capabilities, however, data mining must be performed at the same data level as where the resulting model will operate (as identified in the top row of the table).

Data mining will likely be useful only when applied to a limited subset of data products. This is because there will be known strong correlations (i.e., information redundancy) across similar products (say, varying only in resolution), or similar science parameters (e.g., radiance at 412 nm and 443 nm). Good practice in data mining involves discarding dimensions/fields that are clearly of little interest with regard to the data mining goal. It would not be unusual to use only one in ten available fields, and this ratio is likely to be much higher when dealing with science data.

Sampling is likely to be used extensively in data mining, in keeping with common practice. When the goal is to build a descriptive statistical model, only a representative subset of the data is needed; additional data provide no additional information. Clustering, on the other hand, may more typically be applied to the full set of data, unless the goal is only to partially label a representative set of data used by a model building process. Statistical summaries are also likely to be used, since most of the needed information is contained in the summary.[10]

As discussed earlier, data mining may require a large number of derived parameters to make certain implicit information explicit to the mining algorithm. Although only a few derived parameters may be of significance, that fact is typically not known ahead of time, so more derived parameters may be required as input than will be contained in the output model.

Most data mining algorithms are very computationally complex, and make aggressive use of heuristics and other mechanisms to try to contain that complexity. This complexity, combined with a potentially large data volume, warrants caution. From the table, it become clear that focusing on a representative set of data can reduce the volume of data to be handled by two orders of magnitude or more. When thus constrained, the total data volume is well within the range of feasibility.

[10] For example, average monthly SST as used in [13].

1.4.5 Dynamic Feedback Loops

In the conceptual specimen architecture, this capability is fed by information from event detection or data quality assessment. As such, it shares the characteristics of those capabilities relative to data volume considerations. One difference is that some algorithms may be specifically designed to operate on L0 data, and perhaps collocated with the sensor, to provide low-latency feedback to sensors. In such cases, the filters would most likely be applied to the entire data stream, or at least a substantial portion of it.

1.5 Illustrative Scenario

Fire prediction is one scenario that can help illustrate this discussion of data volumes. Two stages of data mining are envisioned: one to determine precursors of large-scale fire events, and another to detect those precursors in the stream of real-time data.

Using the summary analysis table above as a guide, we can estimate the data volume that would need to be examined to achieve a fire prediction capability.

	Precursor Determination	Precursor Detection
Data Product Levels (Input)	• L3	• L3
Other Subsets	• Locations of known fires (10,000) • Additional locations without fires (100,000) • Previous five years (1825 days) • Global • Selected parameters (10)	• Land Only (30%) • Real-time • Global (500M km^2) • Selected parameters (10)
Sampling	• 100% (Locations with fires) • Nearby locations (without fires)	• None.
Statistical Summaries	• None	• None.
Derived Data	• Time, location, & weather attributes (10)	• Time, location, & weather attributes (10)
Aggregate Input Data Volume	• 0.0002% • (20 GB)	• 0.2% • (7 TB/year)
Algorithm Computational Complexity	• High	•

Focusing on land conditions (e.g., vegetation index, soil moisture) and closely related atmospheric data (e.g., recent precipitation and surface winds) as potential precursors suggests that high-resolution L3 data products might be most appropriate to analyze, since we will need to assess the relationship between these attributes at specific fixed locations (i.e., where a fire occurred).

An appropriate subset of data would include locations of known fires (for the positive case) and a sample of locations without fires (for the negative case). A fairly lengthy prior time period, say five years, might be considered relevant. The subset would also include perhaps on the order of

ten parameters representing various land conditions and related atmospheric data. Assuming we examine something on the order of (at most) 10,000 fires and ten times that number of locations where no fire occurred yields about 200 million observations (110,000 locations x 1825 days).

For fire prediction using precursor detection, an appropriate subset of the data would include high-resolution (say, 1km), real-time data over land. If we are only interested in fire prediction for parochial purposes (e.g., redeployment of U.S. fire fighting resources), the geographic area might be further reduced. Otherwise, no other subsetting opportunities are obvious.

There are no obvious opportunities to use statistical summaries, so we assume no reduction there. It is likely, however, that a fair amount of derived data might be useful, including attributes about time (e.g., season, day of week), location (e.g., proximity to populated areas), and other ancillary data (e.g., thunderstorm presence). Assuming on the order of 10 such parameters in addition to the ten science parameters, and assuming each parameter can be represented in two bytes yields a total data volume of roughly 8 GB, or 0.0002% of a 4 PB archive for precursor determination (200M x 20 x 2). Similarly, we could expect roughly 20 GB for each observation period for precursor detection (500M x 20 x 2). This amounts to roughly 7 TB per year for daily observations, the equivalent of 0.2% of a 4 PB archive. Fortunately, in this scenario, the data volumes are low when the computational complexity is high and, conversely, the computational complexity is low when the data volumes are high. So the data volume problem is tractable, though challenging.

1.6 Conclusions

The data volumes that must be handled to realize the envisioned capabilities of an intelligent data archive can be substantial. However, for those capabilities where the volume is largest (virtual data products, autonomous event detection, automated data quality assessment, and dynamic feedback loops), the relative complexity of the associated algorithms is low. As a result, realizing these capabilities could require no more additional computing capacity than is available today (i.e., a 100% increase). But these capabilities all depend on large scale data mining, which has a moderate to high computational complexity. Fortunately, the total data volume that would have to be mined to provide meaningful results is far less than the total archive volume, ranging perhaps from 2 GB to 20 TB of a 4 PB archive. The volume to be mined can be controlled by focusing L3 data products, a subset of the available products at that level, and global daily averages. The volume can further be constrained by using only representative samples when the goal is to infer a statistical model from the data. Using moderate complexity algorithms against this constrained data volume is clearly challenging but also feasible. As additional computational capacity becomes available, more data can be examined by more algorithms. There is only one obvious pathological case that probably will remain infeasible for the foreseeable future: running unsupervised classification on the entire data stream, or the stream at one data product level. We estimate that these algorithms could require 50 times the computational capacity of the current science algorithms, so the cost of running them on a substantial portion of the data stream would be prohibitive until computing costs fall by perhaps two orders of magnitude.

1.7 References

1. Clausen, Mark and Christopher Lynnes, July, 2003. *Virtual Data Products in an Intelligent Archive*, White Paper prepared for the Intelligent Data Understanding program, 15 p. Link

2. Harberts, Robert, L. Roelofs, H. K. Ramapriyan, G. McConaughy, C. Lynnes, K. McDonald and S. Kempler, 2003. *Intelligent Archive Visionary Use Case: Advanced Weather Forecast Scenario*, White Paper prepared for the Intelligent Data Understanding program, 15 p. Link

3. Harberts, Robert, L. Roelofs, H. K. Ramapriyan, G. McConaughy, C. Lynnes, K. McDonald and S. Kempler, September, 2003. *Intelligent Archive Visionary Use Case: Precision Agriculture Scenario*, White Paper prepared for the Intelligent Data Understanding program, 19 p. Link

4. Harberts, Robert, L. Roelofs, H. K. Ramapriyan, G. McConaughy, C. Lynnes, K. McDonald and S. Kempler, Decemberr, 2003. *Intelligent Archive Visionary Use Case: Virtual Observatories*, White Paper prepared for the Intelligent Data Understanding program, 21 p. Link

5. Isaac, David and Christopher Lynnes, January, 2003. *Automated Data Quality Assessment in the Intelligent Archive*, White Paper prepared for the Intelligent Data Understanding program, 17 p. Link

6. Lynnes, Christopher, July, 2003. *Automated Data Discovery and Usage*, White Paper prepared for the Intelligent Data Understanding program, 13 p. Link

7. McConaughy, Gail and Kenneth McDonald, September, 2003. *Moving from Data and Information Systems to Knowledge Building Systems: Issues of Scale and Other Research Challenges*, White Paper prepared for the Intelligent Data Understanding program, 21 p. Link

8. Morse, H. Stephen, David Isaac, and Christopher Lynnes, January, 2003. *Optimizing Performance in Intelligent Archives*, White Paper prepared for the Intelligent Data Understanding program, 34 p. Link

9. Ramapriyan, H. K., Gail McConaughy, Christopher Lynnes, Steve Kempler, Ken McDonald Bob Harberts, Larry Roelofs and Paul Baker, August, 2002. *Conceptual Study of Intelligent Archives of the Future*, Report prepared for the Intelligent Data Understanding program, 39 p. Link

10. Ed Masuoka, *MODAPS Reprocessing Collections 4 and 5*. MODIS Science Team Meeting, July 13, 2004. Link

11. Christopher Justice et al., *The Moderate Resolution Imaging Spectroradiometer (MODIS): Land Remote Sensing for Global Change Research*. IEEE Transactions on Geoscience and Remote Sensing, Vol. 36, No. 4, July 1998. Link

12. James Tilton, *Hierarchical Image Segmentation*. Link

13. Michael Steinbach, et al., *Discovery of climate indices using clustering*. Conference on Knowledge Discovery in Data, Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C., 2003, pp. 446 – 455.

14. I. A. Galkin et al., *Processing Radio Plasma Imager Plasmagrams Utilizing Hierarchical Segmentation*. (Link)